A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation.

Authors

Muhammad Faisal, *Senior Research Fellow in Medical Statistics*
Faculty of Health Studies, University of Bradford, Bradford, UK
Bradford Institute for Health Research
E-mail: M.Faisal1@bradford.ac.uk

Andy Scally, Medical Statistician
Faculty of Health Studies, University of Bradford, Bradford, UK
Bradford Institute for Health Research
E-mail: A.J.Scally@Bradford.ac.uk

Robin Howes, *Operational Manager for Electronic Patient Records*
Department of Strategy & Planning
Northern Lincolnshire and Goole Hospitals
E-mail: robin.howes@nhs.net

Kevin Beatson, *Development Manager*
York Teaching Hospital NHS Foundation Trust
E-mail: Kevin.Beatson@York.NHS.uk

Donald Richardson, *Consultant Renal Physician*
Department of Renal Medicine, York Teaching Hospital NHS Foundation Trust Hospital
E-mail: drichardson@doctors.org.uk

Mohammed A Mohammed
Professor of Healthcare Quality & Effectiveness
Faculty of Health Studies, University of Bradford, Bradford, UK
Deputy Director of the Bradford Institute for Health Research
Academic Director to the Yorkshire & Humberside Academic Health Sciences Network
E-mail: M.A.Mohammed5@Bradford.ac.uk

Correspondence to: Mohammed A Mohammed

ABSTRACT

We compare the performance of logistic regression with several alternative machine learning methods to estimate the risk of death for patients following an emergency admission to hospital based on the patients' first blood test results and physiological measurements using an external validation approach. We trained and tested each model using data from one hospital (n=24696) and compared the performance of these models in data from another hospital (n=13477). We used two performance measures – the calibration slope and area under the curve (AUC). The logistic model performed reasonably well – calibration slope 0.90, AUC 0.847 compared to the other machine learning methods. Given the complexity of choosing tuning parameters of these methods, the performance of logistic regression with transformations for in-hospital mortality prediction was competitive with the best performing alternative machine learning methods with no evidence of overfitting.

Key words: statistical modelling, classification and prediction, computer intensive methods, modelling healthcare services, electronic health records, databases and data mining

Introduction

Several predictive models are in widespread use to predict the risk of death for patients in hospital. Prominent examples include Acute Physiology and Chronic Health Evaluation (APACHE II) [1], Mortality Probability Model (MPM II) [2] and the Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM) [3].

The development of such risk prediction models is less than straight forward, involving a number of important modelling choices [4] which require consideration of candidate covariates (eg the patients' age, gender, comorbidities), the linearity or otherwise of covariates, interaction effects and the choice of model (eg logistic regression) [5]. Model development is usually guided by a number of model diagnostics and performance statistics such as model calibration and model discrimination [6].

Our motivation stems from attempting to predict the risk of dying for acutely ill patients who are admitted to hospital as unplanned or emergency medical admissions [7]. The response variable is whether the patient died in hospital (yes/no) and the covariate set is based on previous work [8] which identified the patients' routine blood tests (seven blood tests, see later) and National Early Warning Score (NEWS)[9], (see later) as appropriate predictor variables along with the patients age (years) and gender (male/female).

A fundamental issue is choice of model. Here we consider the more traditional approach (which tend to produce models which are more understandable by humans) versus more modern machine learning approaches:- (1) logistic regression without transformations of continuous covariates (LOGIT), (2) logistic regression with transformations of continuous covariates (LOGIT†), (3) logistic regression with multivariable fractional polynomials (MFP) [10], (4) logistic regression with restricted cubic splines (RCS) for continuous covariates [5], (5) recursive partitioning and regression trees (RPART) [11], (6) random forest (RF) [12], (7) generalized boosted regression modelling (GBM) [13], (8) support vector machine (SVM) [14], (9) neural network (NNET) [14].

The rationale for investigating these alternative approaches is as follows. Logistic regression is widely used in medical applications and the model coefficients can be interpreted as odds ratios (and using a modified approach as risk ratios) which are clinically meaningful [4]. For the logistic model, covariates can be included on the untransformed scale, with transformation and with/without the aid of restricted cubic splines (RCS) (which are advocated for use with continuous covariates [5,15]) and MFP (which has also has been advocated for continuous covariates [16]). Furthermore, modern statistical machine learning methods have been advocated by several authors [17–25], including decision trees, boosted models, support vector machine, and neural networks.

For this paper, we consider logistic regression with/without transformation as being the more traditional approach and the use of RCS, MFP, RF, RPART, GBM and SVM as the more modern computationally intensive approaches.

Our aim is to compare the above modelling strategies and identify the model with the best performance statistics using external model validation to assess the performance of these models in terms of calibration and discrimination. The use of external validation to make these comparisons has become an important methodological development [6,26].

As prediction models inform patients and carers about prognosis, it is essential that predictions should be well calibrated [6]. Whilst the interest in the development and validation prediction models in clinical setting is growing, only a quarter of the studies reported prediction models with internal and external validation [27,28]. Usually internal validation is done by splitting the development data into training and testing sets however cross-validation and bootstrapping can also be used [29]. External validation aims to address the performance of a model in patients from a different but possibly related setting, and it is a key step before disseminating prediction model in clinical setting [26,30].

For discrimination, we use the area under the receiver-operator curve (AUC) or concordance (c)-statistic. The AUC is the probability that the model will predict a higher risk of death for a randomly selected patient who died, compared to a randomly selected patient who survived.

Calibration is the relationship between the observed and predicted risk of death and can be usefully seen on a scatter plot (y-axis observed risk, x-axis predicted risk). Perfect predictions should be on the 45° line. The intercept (a) and slope (b) of this line gives an assessment of '*calibration-in-the-large'* [6]. At model development, $a = 0$ and $b = 1$, but at external validation, calibration-in-the-large problems are indicated if *a* is not 0 and if *b* is more/less than 1 as this reflects problems of under/over prediction. Specifically, for each modelling strategy, we determined the AUC or c-statistic, the scaled Brier score, the Hosmer-Lemeshow deciles of risk goodness of fit test (HL).

Materials and Methods

Data set

Our cohorts of emergency admissions are from two acute hospitals which are approximately 100 kilometres apart in the Yorkshire & Humberside region of England– the Diana, Princess of Wales Hospital (managed by the Northern Lincolnshire and Goole NHS Foundation Trust (NLAG)), and York Hospital (managed by York Teaching Hospitals NHS Foundation Trust). All adult (age>16 years) emergency admissions during the year 2014 (i.e., 1st January 2014 to 31st December 2014) were included. We obtained the following information for each admission: the patients' age, gender, and discharge status (alive/dead). We considered admissions, which had no missing data. We excluded 5137 (17%) admissions for NLAG Hospital and 4267 (24%) admission for York Hospital, with incomplete data (albumin and creatinine test results were the most frequent missing data) (**Table 1**). The covariates set was:- age (years), gender (male/female), albumin (g/L), creatinine (µmol/L), haemoglobin (g/dL), potassium (mmol/L), sodium (mmol/L), white cell count (10^9 cells/L), urea (mmol/L), and national early warning score (NEWS). The NEWS ranged from 0 (indicating the lowest severity of illness) to 19 (the maximum NEWS value possible is 20).

Statistical Analysis

We started with an exploratory analysis of the NEWS and the blood test results. We truncated extreme observations of blood test results (very high (>99.9% centile) or very low (<0.1% centile) to moderate the noise of outliers in the modelling process. We have excluded the incomplete data as follows: 17% (5138/29834) for NLAG hospital and 24% (4267/17744) for York hospital. We produced scatter plots showing the relationship between mortality and continuous covariates (grouped into deciles). We modelled the risk of death using the same set of covariates:- age, gender, albumin, creatinine, haemoglobin, potassium, sodium, white cell count, urea, and NEWS.

We used the *qladder* function (Stata version 13), which displays the quantiles of transformed variable against the quantiles of a normal distribution according to the ladder powers $\left(x^3, x^2, x^1, x, \sqrt{x}, \log(x), x^{-1}, x^{-2}, x^{-3}\right)$ for each variable x. We randomly divided our development data (NLAG Hospital) into a training set (70%, n = 17288) and a testing set (30%, n = 7408) for internal model validation [4]. We further validated these models on an external validation dataset from York hospital. Three commonly used performance measures were used to assess model performance:- Hosmer–Lemeshow (HL) test, scaled Brier score, and area under the ROC curve (AUC) [6]. The 95% confidence interval (95%CI) for the c-statistic was derived using DeLong's method as implemented in the pROC library [31]. Discrimination relates to how well a model can separate, (or discriminate between), those who died and those who did not. Calibration relates to the agreement between observed mortality and predicted risk. Overall statistical performance was assessed using the scaled Brier score which incorporates both discrimination and calibration [4]. The Brier score is the squared difference between actual outcomes and predicted risk of death, scaled by the maximum Brier score such that the scaled Brier score ranges from 0–100%. Higher values indicate superior models.

These analyses were undertaken in R [32]. We used default tuning parameters for MFP in R packages mfp [33] but RCS with three knots in R packages rms [34]. We used the caret R

package [35] for machine learning algorithms (RPART, RF, GBM, SVM, NNET) and optimised their tuning parameters using AUC as a loss function for a (five times) repeated 10-fold cross validation method (see supplementary material). We used a linear kernel with two parameters (i.e., cost and gamma) for the SVM method.

Ethical Approval

Although this type of study does not require ethical approval because it meets the exemption criteria ("Research limited to secondary use of information previously collected in the course of normal care (without an intention to use it for research at the time of collection), provided that the patients or service users are not identifiable to the research team in carrying out the research.[36])" we obtained ethical approval for the main research project of which this is a sub study from Yorkshire & The Humber - Leeds West Research Ethics Committee (reference number 15/YH/0348).

Results

There were 24696 emergency admissions for development data (NLAG Hospital) and 13477 for validation data (York Hospital). We further divide the development data into training set (70%, n=17288) and testing set (30%, n=7408). For both hospitals, we have 12-months data where patient discharges were from 1st January 2014 to 31st December 2014. Descriptive statistics for the covariates are shown in Table 1. The risk of death in NLAG was 4.7% (1159/24696) compared with 6.5% (876/13477) in York hospital. Patients in NLAG hospital has a mean age of 63.1 years compared with 68.3 years in York hospital and a lower NEWS (1.9 (NLAG) compared to York hospital (2.6)

Figure 1 shows box plots of each covariate with respect to patient discharge status (Alive/Dead) in NLAG hospital. In general, patients who died were older, had higher NEWS, lower albumin, higher creatinine, lower haemoglobin, higher potassium, higher urea, higher white cell counts and lower sodium levels.

Figure 2 shows that the relationship between the continuous covariates and mortality in NLAG hospital is generally non-linear. Using quantile-quantile (*qq*) plots, we arrived at the following transformations: $(\text{creatinine})^{-0.5}$, $\log(\text{potassium})$, $\log(\text{sodium})$, $\log(\text{white cell count})$, $\log(\text{urea})$.

Statistical Modelling Results

We predicted the risk of in-hospital mortality using the following modelling approaches – LOGIT (no transformations), LOGIT† (with transformations), MFP, RCS, RPART, RF, GBM, SVM, NNET. The model performance statistics are shown in Table 2 and plotted in Figure 3.

In the training phase the AUC ranged from 0.87 to 1. RF had a perfect AUC (1) which is a reflection of the overfitting that usually occurs when RF trees are grown to the maximum size in training datasets using the default (and recommended) settings. GBM had the highest AUC (0.905). RPART had the lowest AUC (0.869). The other five methods (LOGIT, LOGIT†, MFP, NNET, SVM) had AUCs that ranged from 0.883 to 0.884. In the training phase RF had the highest Brier score (0.884 which is also due to over fitting) followed by GBM and RPART (0.244, 0.261). The remaining six methods (LOGIT, LOGIT†, RCS, MFP, SVM, NNET) had Brier scores that ranged from 0.158 to 0.164.

In the testing phase all methods had a reduction in their AUC (range: 0.814 to 0.872) and Brier scores (range: 0.025 to 0.164). RPART now had the lowest AUC (0.814) followed by RF (0.857). The remaining seven methods (LOGIT, LOGIT†, RCS, MFP, RF, NNET, GBM, SVM) had very similar AUC that ranged from 0.871 to 0.872. In the testing phase, RPART had the lowest Brier score (0.025) followed by RCS (0.080). The highest Brier Score was seen in RF (0.164). The remaining six methods (LOGIT, LOGIT†, MFP, GBM, NNET, SVM) had Brier scores that ranged from 0.111 to 0.131.

In the external validation phase all methods had a reduction in their AUC (range: 0.785 to 0.851) and Brier Scores (0.048 to 0.149). The highest AUC (0.851) and Brier Score (0.149)

was seen in the LOGIT† model. The lowest AUC was seen in RPART (0.785) followed by RF (0.804). The remaining models (LOGIT, RCS, MFP, GBM, SVM, NNET) had AUC that ranged from 0.847 to 0.851. The lowest Brier Score was seen in RPART (0.048) followed by RF (0.119), whilst the remaining methods (LOGIT, LOGIT†, RCS, MFP, GBM, NNET, SVM) had Brier Scores which ranged from 0.135 to 0.149.

The external validation calibration slope (Figure 3, lower panel) ranged from 0.70 to 0.99, with RPART having the lowest value which showed considerable over-fitting (slope<1). Three methods (RCS, LOGIT† and GBM) had a 95%CI which included 1. GBM had an external validation slope nearest to one, 0.99.

The LOGIT model without transformations performed reasonably well in the external validation phase – AUC (0.847), Brier Score (0.139) and Slope (0.90). The LOGIT† also performed well – AUC (0.847), Brier Score (0.149) and Slope (0.92). The RCS and MFP models also had similar AUC (0.85) and Brier Scores (0.138 and 0.148 respectively). The RCS slope was higher 0.93 with a wider 95%CI that included one, whilst MFP had a slope of 0.91 and a narrower 95%CI which did not include one. As the sample size is the same for all methods, the source of variability in the width of the confidence intervals is linear predictors from each method and identifies RCS as having the widest confidence intervals for their estimates of the external validation slope.

Discussion

Using a high quality electronically collected data set with large sample sizes and non-linear relationships between covariates and mortality, we examined the performance of nine methods for predicting the risk of in hospital mortality by developing the model in one hospital and externally validating it in another hospital. This approach to model testing is infrequent [27,28] but methodologically more rigorous than simply considering internal validation [26].

We did not find any consistent evidence to suggest that modern machine learning approaches (RPART, RF, GBM, SVM, NNET) were superior to more conventional statistical modelling methods based on the logistic regression model. Whilst there was no clear overall winner, GBM and LOGIT† exhibited the best overall performance. However, we did find that several methods (RPART, RF) exhibited sufficiently poor performance in the external validation phase to undermine their use. Furthermore, given the complexity of choosing tuning parameters of the alternative machine learning methods the logistic regression with transformations has good performance characteristics and is relatively less complex.

Although a few studies have used external validation as a benchmark for machine learning and logistic regression methods in following areas: detecting prostate cancer [37,38], on simulated data [39], predicting mortality risk after acute ischemic stroke [40] and predicting mortality risk after brain injury [23,41], we predicted the risk of in-hospital mortality in acutely ill medical admissions. Our findings are consistent with recently published study on predicting the risk of mortality after traumatic brain injury [41]. As they found logistic regression performs just as well as modern machine learning methods. A key reason for this may be that nonlinear and non-additive signals are not strong enough to make modern machine learning methods advantageous.

Whilst the extent to which our findings are generalisable is not clear, we suggest that candidate models should include a simple logistic model as a benchmark for comparison with other more sophisticated models and that external validation (not internal validation) be

used to compare and contrast model performance. Furthermore, the use of the AUC alone as a summary performance measure is limited and not necessarily a good discriminator between models. The Brier score, which combines calibration and discrimination and the external calibration slope are also useful performance characteristics which merit consideration when comparing models.

Conclusion: Given the complexity of choosing tuning parameters of the modern machine learning methods considered above, the performance of logistic regression with transformations for in-hospital mortality prediction was competitive with the best performing alternative machine learning methods with no evidence of overfitting. The use of RPART and RF in our data is not supported. Our models were developed (using training and testing datasets) in one hospital and validated in a second (different) hospital within the region which increases the likelihood of generalisability to other hospitals. Having established the validity of the logistic regression modelling approach, we plan to evaluate its use in routine clinical practice to see if it can support clinical decision making to enhance the quality of care.

Funding

This research was supported by the Health Foundation. The Health Foundation is an independent charity working to improve the quality of health care in the UK.

This research was supported by the National Institute for Health Research (NIHR) Yorkshire and Humberside Patient Safety Translational Research Centre (NIHR YHPSTRC). The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Contributorship

MAM & MF had the original idea for this work and undertook the statistical analyses. RH and KB extracted the necessary data frames. DR gave a clinical perspective. AS contributed in study design

and interpretation of results. MAM and MF wrote the first draft of this paper and all authors subsequently assisted in redrafting and have approved the final version.

Competing Interests

The authors declare no conflicts of interest.

References

1 Knaus WA, Draper EA, Wagner DP, *et al.* APACHE II: a severity of disease classification system. *Crit Care Med* 1985;**13**:818–29.http://www.ncbi.nlm.nih.gov/pubmed/3928249 (accessed 19 Feb 2015).

2 Lemeshow S, Teres D, Klar J, *et al.* Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993;**270**:2478–86. doi:10.1001/jama.270.20.2478

3 Neary WD, Heather BP, Earnshaw JJ. The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM). *Br J Surg* 2003;**90**:157–65. doi:10.1002/bjs.4041

4 Steyerberg EW. *Clinical Prediction Models. A practical approach to development, validation and updating.* Springer 2008.

5 Harrell FE. *Regression Modeling Strategies : With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer New York 2001.

6 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–38. doi:10.1097/EDE.0b013e3181c30fb2

7 Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013;**3**:e003482–e003482. doi:10.1136/bmjopen-2013-003482

8 Prytherch DR, Sirl JS, Schmidt P, *et al.* The use of routine laboratory data to predict in-hospital death in medical admissions. *Resuscitation* 2005;**66**:203–7. doi:10.1016/j.resuscitation.2005.02.011

9 Physicians RC of. National. Early Warning Score (NEWS): Standardising the assessment of acuteillness severity in the NHS https://www.rcplondon.ac.uk/sites/default/files/documents/national-early-warning-score-standardising-assessment-acute-illness-severity-nhs.pdf. 2012.

10 Royston P, Sauerbrei W. *Fractional Polynomials for One Variable Multivariable Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. John Wiley 2008.

11 Breiman L, Friedman J, Stone JC, *et al. Classification and regression trees*. 1st ed. Wadsworth: : Wadsworth International Group 1984.

12 Breiman L. Random Forests. *Mach Learn*;**45**:5–32. doi:10.1023/A:1010933404324

13 Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001;**29**:1189–232.http://cat.inist.fr/?aModele=afficheN&cpsidt=13468948 (accessed 18 Nov 2015).

14 Hastie T, Tibshirani R, Friedman HJ. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: : Springer Science Business Media, LLC 2009.

15 Marrie RA, Dawson N V, Garland A. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *J Clin Epidemiol* 2009;**62**:511–7.e1.

doi:10.1016/j.jclinepi.2008.05.015

16 Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials. *Methods Inf Med* 2005;**44**:561–71.http://www.ncbi.nlm.nih.gov/pubmed/16342923 (accessed 25 May 2015).

17 Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthc Inform Res* 2011;**17**:232–43. doi:10.4258/hir.2011.17.4.232

18 Scott HF, Colborn K. Machine learning for predicting sepsis in-hospital mortality: an important start. *Acad Emerg Med* Published Online First: May 2016. doi:10.1111/acem.13009

19 Churpek MM, Yuen TC, Winslow C, *et al.* Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Crit Care Med* 2016;**44**:368–74. doi:10.1097/CCM.0000000000001571

20 Badriyah T, Briggs JS, Prytherch DR. Decision Trees for Predicting Risk of Mortality using Routinely Collected Data. 2012;**6**:660–3.

21 Wang G, Lam K-M, Deng Z, *et al.* Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. *Comput Biol Med* 2015;**63**:124–32. doi:10.1016/j.compbiomed.2015.05.015

22 Motwani M, Dey D, Berman DS, *et al.* Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 2016;**52**:468–76. doi:10.1093/eurheartj/ehw188

23 Stylianou N, Akbarov A, Kontopantelis E, *et al.* Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. *Burns* 2015;**41**:925–34. doi:10.1016/j.burns.2015.03.016

24 Colombet I, Ruelland A, Chatellier G, *et al.* Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp* 2000;:156–60.http://www.ncbi.nlm.nih.gov/pubmed/11079864 (accessed 27 Jun 2016).

25 Ross EG, Shah NH, Dalman RL, *et al.* The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J Vasc Surg* Published Online First: 2016. doi:10.1016/j.jvs.2016.04.026

26 Bleeker S., Moll H., Steyerberg E., *et al.* External validation is necessary in prediction research:: A clinical example. *J Clin Epidemiol* 2003;**56**:826–32. doi:10.1016/S0895-4356(03)00207-5

27 Collins GS, de Groot JA, Dutton S, *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;**14**:40. doi:10.1186/1471-2288-14-40

28 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al.* External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;**68**:25–34. doi:10.1016/j.jclinepi.2014.09.007

29 Steyerberg EW, Harrell FE, Borsboom GJJ., *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;**54**:774–81. doi:10.1016/S0895-4356(01)00341-9

30 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515–24.http://www.ncbi.nlm.nih.gov/pubmed/10075620 (accessed 27 Jun 2016).

31 Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77. doi:10.1186/1471-2105-12-77

32 R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing http://www.r-project.org/. 2015.

33 Ambler G, Benner A. mfp: Multivariable fractional polynomials. 2015.http://cran.r-

project.org/package=mfp

34 Harrell FE. rms: Regression Modeling Strategies http://cran.r-project.org/package=rms. 2015.

35 Kuhn. M, Wing J, Weston S, *et al.* caret: Classification and Regression Training. 2015.http://cran.r-project.org/package=caret

36 NHS Health Research Authority. Governance Arrangements for Research Ethics Committees. http://www.hra.nhs.uk/resources/research-legislation-and-governance/governance-arrangements-for-research-ethics-committees/ (accessed 10 Aug 2017).

37 Ecke TH, Hallmann S, Koch S, *et al.* External Validation of an Artificial Neural Network and Two Nomograms for Prostate Cancer Detection. *ISRN Urol* 2012;**2012**:1–6. doi:10.5402/2012/643181

38 Nieboer D, Vergouwe Y, Roobol MJ, *et al.* Nonlinear modeling was applied thoughtfully for risk prediction: the Prostate Biopsy Collaborative Group. *J Clin Epidemiol* 2015;**68**:426–34. doi:10.1016/j.jclinepi.2014.11.022

39 Terrin N, Schmid CH, Griffith JL, *et al.* External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. *J Clin Epidemiol* 2003;**56**:721–9. doi:10.1016/S0895-4356(03)00120-3

40 König IR, Malley JD, Weimar C, *et al.* Practical experiences on the necessity of external validation. *Stat Med* 2007;**26**:5499–511. doi:10.1002/sim.3069

41 van der Ploeg T, Nieboer D, Steyerberg EW. Modern modelling techniques had limited external validity in predicting mortality from traumatic brain injury. *J Clin Epidemiol* Published Online First: 14 March 2016. doi:10.1016/j.jclinepi.2016.03.002

Table 1 Characteristics of the emergency admissions in the two hospitals

Characteristic	Development data (NLAG Hospital)	Validation data (York Hospital)
N	24696	13477
Died	1159 (4.7%)	876 (6.5%)
Male	11571 (46.9%)	6413 (47.6%)
Mean Age [years] (SD)	63.1 (21.1)	68.3 (19.2)
Mean NEWS [1-19] (SD)	1.9 (2.1)	2.6 (2.6)
Mean Albumin [g/L] (SD)	34 (6.2)	38 (5.8)
Mean Creatinine [umol/L] (SD)	100.1 (75.2)	104 (93.7)
Mean Haemoglobin [g/dL] (SD)	128.8 (21.7)	125.1 (22.1)
Mean Potassium [mmol/L] (SD)	4.1 (0.6)	4.3 (0.6)
Mean Sodium [mmol/L] (SD)	137 (4.7)	136.7 (4.7)
Mean White cell count [10^9 cells/L] (SD)	9.8 (5.1)	10.3 (7.2)
Mean Urea [mmol/L] (SD)	7.3 (5.7)	8.2 (6.1)

Table 2: Model performance statistics using regression methods for all models (LOGIT, LOGIT†, MFP, RCS, RPART, RF, GBM, SVM, NNET).

Model	Split	HL Chi-squared (df=8)	HL.p	Brier	AUC [95% CI]	Slope [95% CI]
LOGIT	Training	28.0	0.000	0.160	0.8832 [0.8736 – 0.8928]	–
LOGIT	Testing	15.3	0.053	0.111	0.8712 [0.8557– 0.8868]	–
LOGIT	Validation	16.7	0.033	0.139	0.8470 [0.8351– 0.8589]	0.90 [0.85 – 0.96]
LOGIT†	Training	20.1	0.010	0.162	0.8835 [0.8739 – 0.8931]	–
LOGIT†	Testing	14.9	0.061	0.118	0.8714 [0.8558 – 0.8871]	–
LOGIT†	Validation	11.9	0.155	0.149	0.8491 [0.8372 – 0.8610]	0.92 [0.84 – 1.00]
RCS	Training	19.0	0.015	0.169	0.8882 [0.8757 – 0.8947]	–
RCS	Testing	17.5	0.026	0.080	0.8715 [0.8560 – 0.8871]	–
RCS	Validation	27.3	0.001	0.138	0.8476 [0.8356 – 0.8596]	0.93 [0.85 – 1.02]
MFP	Training	19.1	0.014	0.164	0.8850 [0.8756 – 0.8945]	–
MFP	Testing	11.5	0.173	0.115	0.8714 [0.8559 – 0.8870]	–
MFP	Validation	16.1	0.041	0.145	0.8506 [0.8389 – 0.8624]	0.91 [0.86 – 0.97]
RPART	Training	0.0	1.000	0.261	0.8694 [0.8557 – 0.8830]	–
RPART	Testing	47.4	0.000	0.025	0.8137 [0.7898 – 0.8377]	–
RPART	Validation	65.7	0.000	0.048	0.7854 [0.7700 – 0.8007]	0.70 [0.65 – 0.75]
RF	Training	–	–	0.884	1.0000 [1.0000 – 1.0000]	–
RF	Testing	–	–	0.164	0.8569 [0.8397 – 0.8741]	–
RF	Validation	27.9	0.001	0.119	0.8044 [0.7899 – 0.8189]	0.93 [0.86 – 0.99]
GBM	Training	59.3	0.000	0.244	0.9058 [0.8968 – 0.9148]	–
GBM	Testing	30.0	0.000	0.116	0.8719 [0.8563 – 0.8875]	–
GBM	Validation	47.3	0.000	0.142	0.8483 [0.8365 – 0.8601]	0.99 [0.93 – 1.04]
SVM	Training	28.6	0.000	0.158	0.8840 [0.8744 – 0.8936]	–
SVM	Testing	15.9	0.041	0.131	0.8724 [0.8569 – 0.8880]	–
SVM	Validation	113.5	0.000	0.135	0.8470 [0.8351 – 0.8590]	0.89 [0.83 – 0.94]
NNET	Training	11.4	0.182	0.159	0.8842 [0.8747 – 0.8938]	–
NNET	Testing	14.8	0.064	0.123	0.8722 [0.8566 – 0.8877]	–
NNET	Validation	38.3	0.000	0.143	0.8475 [0.8357 – 0.8594]	0.86 [0.80 – 0.91]

†Covariates set is transformed using *qladder* function as follows: $(\text{creatinine})^{-1/2}$, log(potassium), log(sodium), log(white cell count), log(urea).

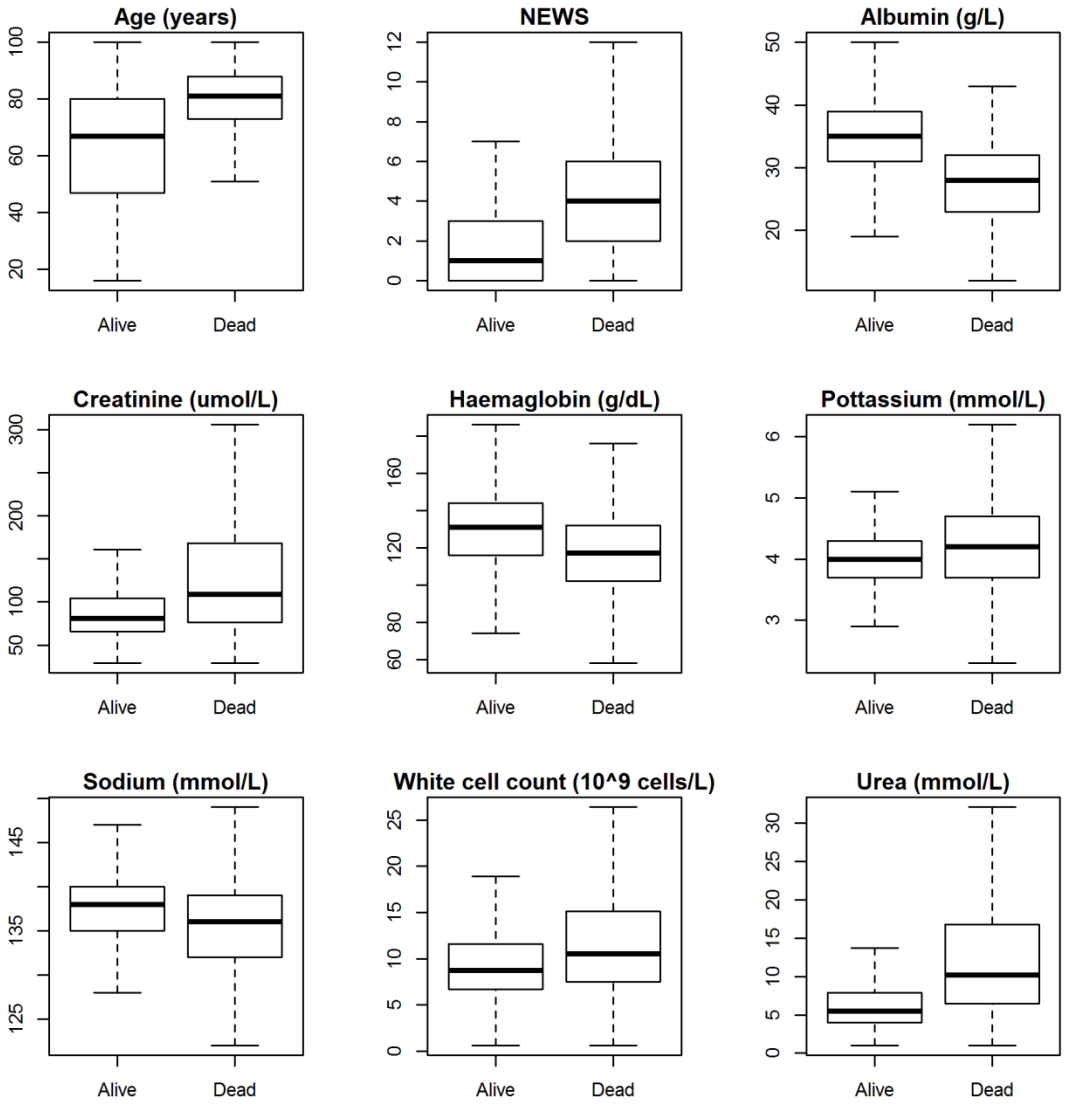

Figure 1: Boxplot without outliers for continuous covariates with respect to patient's discharge status (Alive/Dead) in NLAG hospital

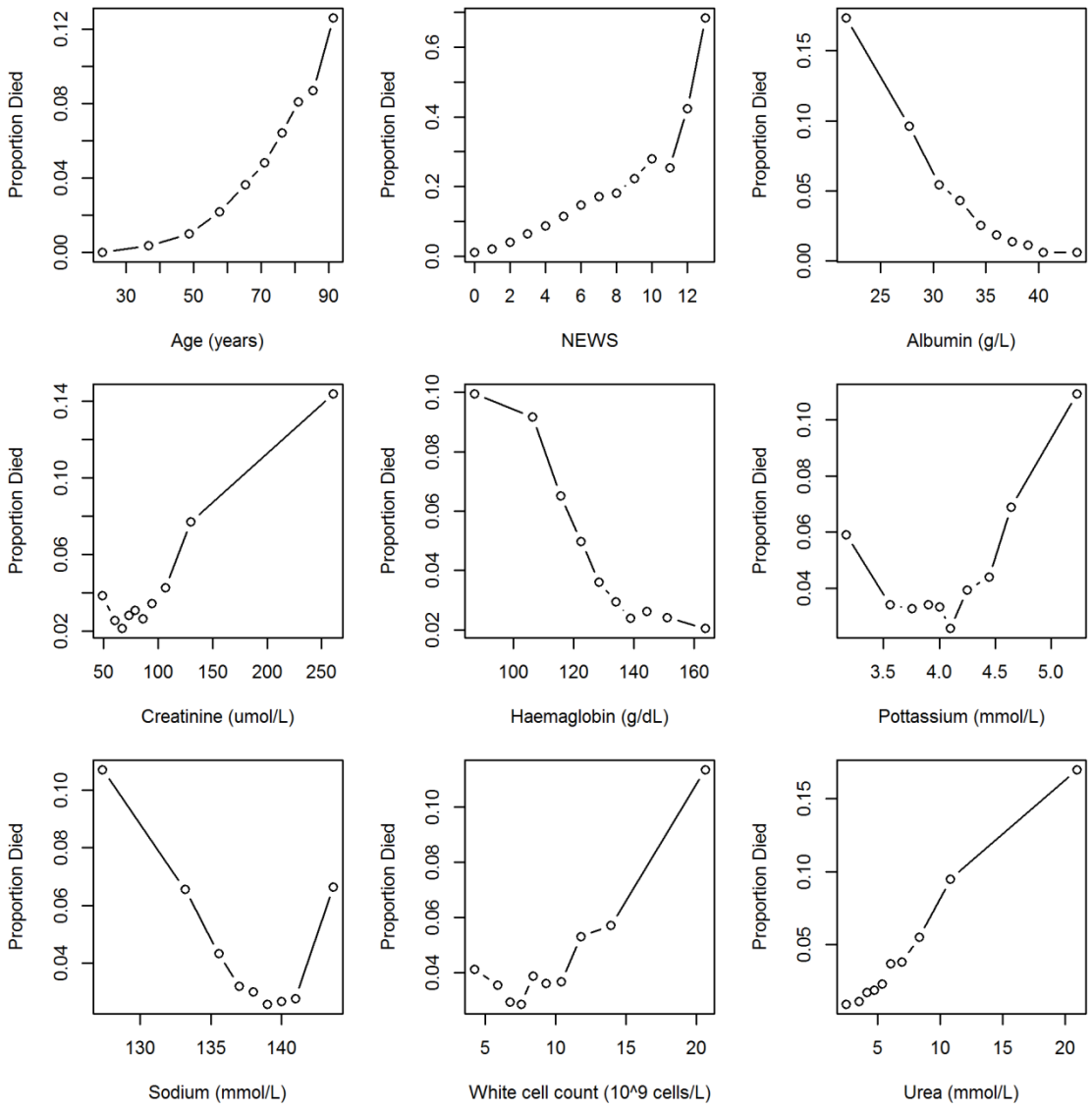

Figure 2: Scatter plots showing the observed risk of death with continuous covariates in NLAG hospital

NB: y-axis range changes in each plot.

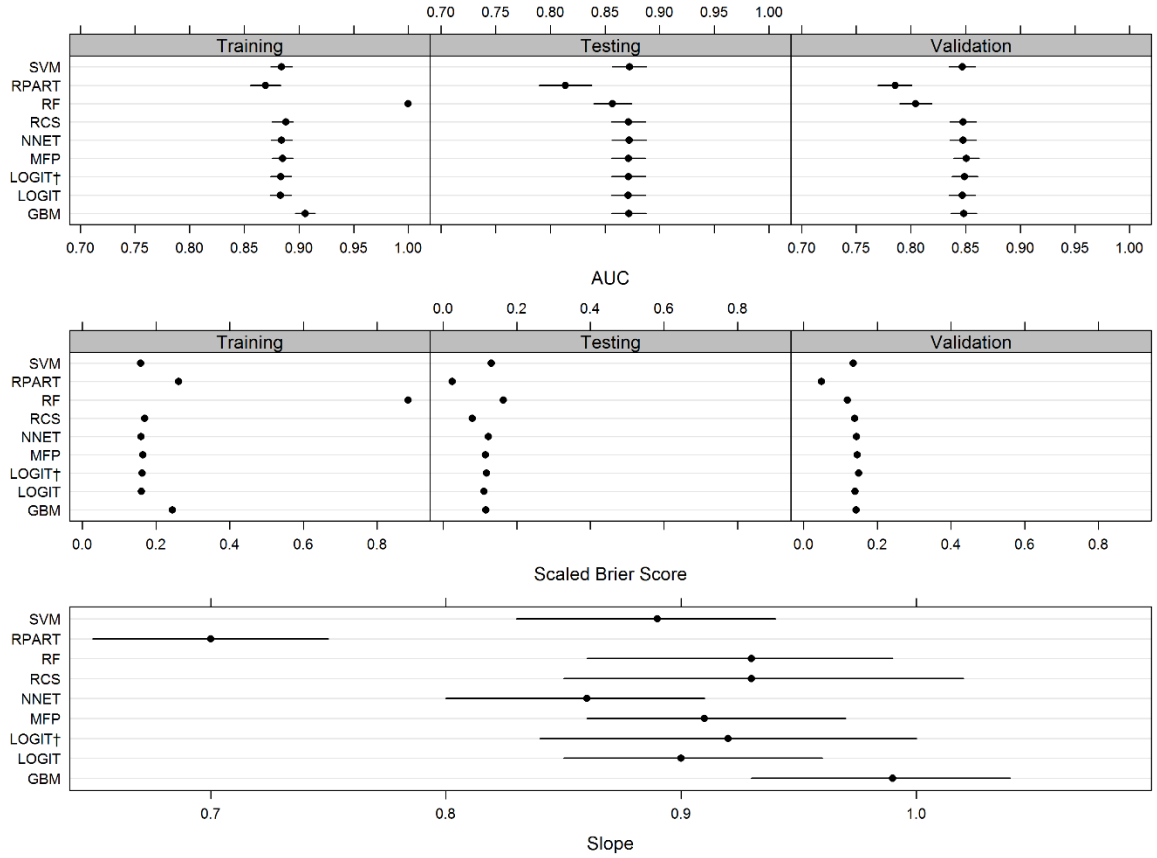

Figure 3: Model performance statistics using regression methods for all models (LOGIT, LOGIT†, MFP, RCS, RPART, RF, GBM, SVM, NNET).
†Covariates set is transformed using *qladder* function as follows: $(\text{creatinine})^{-1/2}$, log(potassium), log(sodium), log(white cell count), log(urea).

Supplementary Material:

Table: Selection of tuning parameters using AUC as loss function in (five times) repeated 10-fold cross validation

Method	Parameter 1	Parameter 2	Parameter 3	Parameter 4
LOGIT	Default	-	-	-
LOGIT†	Default	-	-	-
RCS	Degree=3	-	-	-
MFP	Degree=4	-	-	-
RPART	Cp=0.0001 0 to 0.01 step by 0.0001	-	-	-
RF	TuneLength=10	n.trees=1000	-	-
GBM	n.trees=900 (100 to 2000 step by 100)	Interaction.depth=5 (1,3,5,7)	Shrinkage=0.01 (0.001,0.01,0.1)	n.minobsinnode=16 (1,6,11,16)
SVM	Cost=6 (0.1,0.5,1,2,4,6,8,10)	Gamma=0.1 (0.01,0.05 0.1,0.2,0.5,1)	-	-
NNET	Decay=0.102 (0 to 0.2 step by 0.004)	Size =1 (1,3,5)	-	-

R code:

```
############################
# Modelling - LOGIT
############################

modFormula <- paste("died.train~male +age + NEWS + ALB + CRE + HB  + POT + SOD + WBC + URE ")
modFormula <- as.formula(modFormula)

model_logit = glm(modFormula, data = training, family = "binomial", x=TRUE, y=TRUE)


############################
# Modelling – LOGIT†
############################

modFormulat <- paste("died.train~male +age + NEWS + ALB + sqrt_CRE + HB  + log_POT + SOD + log_WBC +
log_URE")
modFormulat <- as.formula(modFormulat)

model_logit_trans = glm(modFormulat, data = training, family = "binomial", x=TRUE, y=TRUE)

############################
# Modelling - RCS
############################

modFormularcs <- paste("died.train~male +rcs(age,3) + NEWS + rcs(ALB,3) + rcs(CRE,3) + rcs(HB,3)  +
rcs(POT,3) + rcs(SOD,3) + rcs(WBC,3) + rcs(URE,3)")
modFormularcs <- as.formula(modFormularcs)


model_rcs = glm(modFormularcs,family = "binomial",data = training, x=TRUE, y=TRUE)

summary(model_rcs)

############################
# Modelling - MFP
############################

modFormulamfp <- paste("died.train~male +fp(age) + NEWS + fp(ALB) + fp(CRE) + fp(HB)  + fp(POT) +
fp(SOD) + fp(WBC) + fp(URE)")
modFormulamfp <- as.formula(modFormulamfp)

model_mfp = mfp(modFormulamfp,family = "binomial",data = training, x=TRUE, y=TRUE)

############################
# Modelling - RPART
############################
#RPART
set.seed(669)

ctrl <- trainControl(method = "repeatedcv",repeats = 5,number = 10,classProbs = T,summaryFunction =
twoClassSummary)


rpartModel <- train(modFormula,
                    data = training,
                    method = "rpart",
                    tuneGrid =expand.grid(.cp=seq(0,0.01,length=100)),
                    metric="ROC",
                    trControl = ctrl
)

############################
# Modelling - RF
############################
#rf
set.seed(669)
rfModel <- train(modFormula,
                 data = training,
                 method = "rf",
                 tuneLength = 10,
                 ntrees = 1000,
```

```
                importance = TRUE,
                metric="ROC",
                trControl = ctrl
                )

############################
# Modelling - GBM
############################
#gbm
gbmGrid <- expand.grid(.interaction.depth = seq(1,7,by=2), .n.trees = seq(100, 2000, by =
100),.shrinkage = c(0.001,0.01, 0.1),.n.minobsinnode=seq(1,20,by=5))

set.seed(669)
gbmModel <- train(modFormula,
                  data = training,
                  method = "gbm",
                  tuneGrid = gbmGrid,
                  verbose = FALSE,
                  metric="ROC",
                  trControl = ctrl
                  )



############################
# Modelling – SVM
############################
#SVM
ctrl <- trainControl(method = "repeatedcv",repeats = 5,number = 10,classProbs = T,summaryFunction =
twoClassSummary)

set.seed(669)
#creation of weights - also fast for very big datasets
weights <- as.numeric(died.train)-1

for(val in unique(weights)) {weights[weights==val]=1/sum(weights==val)*length(weights)/2} # normalized
to sum to length(samples)

svmModel <- train(modFormula,method = 'svmLinear',
                  maximize = T,
                  weights=weights,
                  tuneGrid=expand.grid(.C=c(0.1,0.5,1,2,4,6,8,10),.sigma=c(0.01,0.05,0.1,0.2,0.5,1)),
                  preProcess = c('center', 'scale'),
                  maxit=10000,
                  data=training,
                  metric="ROC",
                  trControl = ctrl
)


############################
# Modelling - NNET
############################
#NNET
nnetGrid <- expand.grid(.decay = seq(0,0.2,length=50), .size =c(1,3,5))

set.seed(669)
nnetModel <- train(modFormula,
                   data = training,
                   method = "nnet",
                   tuneGrid = nnetGrid,
                   preProc = c("center", "scale"),
                   maxit=10000,
                   trace=F,
                   metric="ROC",
                   trControl = ctrl
)
```